DOUBT: Decoupled Object-level Understanding and Bridging via vMF-based Trustworthiness for Hallucination Detection in MLLMs

Anonymous Authors[1]

Abstract

Multimodal Large Language Models (MLLMs) frequently produce hallucinations (i.e., assertions that contradict the image or facts), undermining reliability in high-risk applications. Existing detection approaches typically feed images and texts jointly and estimate hallucination scores by measuring the consistency of model outputs. However, because the visual module often lags behind the language module in understanding and reasoning, MLLMs can repeatedly produce similar yet incorrect answers, yielding deceptively high measured trustworthiness and therefore missed detections. To address this, we propose a simple yet effective model-agnostic method, dubbed **D**ecoupled **O**bject-level **U**nderstanding and **B**ridging via vMF-based **T**rustworthiness (DOUBT). DOUBT i) elicits richer object-aware responses by decoupling object recognition from relational reasoning via a two-step prompting scheme (Object-level Understanding and Bridging, OUB), and ii) measures reliability with a von Mises–Fisher (vMF)-based trustworthiness metric that is more stable than semantic-entropy metrics under small-sample regimes. Specifically, OUB first prompts the model to list recognized objects, and then conditions chain-of-thought reasoning on those objects to produce object-bridged responses. For trustworthiness estimation, we replace conventional measures with the proposed vMF-based metric, which is robust even under low-sample settings and exhibits smoother behavior than prior techniques. Extensive experiments and ablation studies across multiple benchmarks demonstrate that DOUBT consistently outperforms state-of-the-art baselines, offering a robust and generalizable solution for hallucination detection in MLLMs.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

In recent years, Multimodal Large Language Models (MLLMs) have achieved remarkable progress and shown strong performance on tasks such as visual question answering and multimodal object reasoning (Li et al., 2023a; Wang et al., 2024b). Despite these advances, MLLMs still frequently generate assertions that contradict input images or external facts, commonly referred to as **hallucination** (Bai et al., 2024; Liu et al., 2024c; Huang et al., 2024). Such hallucinations undermine the trustworthiness of MLLM outputs and can pose serious safety risks in high-stakes domains, including medical diagnosis, autonomous systems, legal analysis, scientific discovery, and education (Ji et al., 2023; Bubeck et al., 2023; Nori et al., 2023). As MLLMs are increasingly deployed in real-world applications, hallucination detection has become essential for safe and trustworthy AI (Achiam et al., 2023).

Existing approaches for hallucination detection can be broadly categorized into **white-box** and **black-box** approaches (Huang et al., 2025; Ahadian & Guan, 2025). White-box methods inspect model-internal signals (e.g., attention, activations, gradients) to infer hallucination scores, but require access to the model architecture, which is often infeasible for closed-source models (Dasgupta et al., 2025; Li et al., 2025). Black-box methods, in contrast, operate solely on observable outputs instead of model-internal signals, such as by verifying model outputs against external knowledge bases or by sampling multiple responses to estimate output diversity (Manakul et al., 2023; Li et al., 2024). Although black-box methods are flexible and broadly deployable, most of them often rely on complete external knowledge bases and stable metrics, limiting their robustness in practice. To get rid of such dependence, recent approaches employ MLLMs to generate sufficiently diverse answers for uncertainty estimation in hallucination detection (Chen et al., 2024a), thereby providing a generalizable and easily deployable solution.

Although the uncertainty-based methods (Farquhar et al., 2024) are effective at detecting hallucinations, they face two major challenges in practice. First, MLLMs tend to repeatedly generate semantically similar but incorrect answers when faced with object hallucinations, producing low

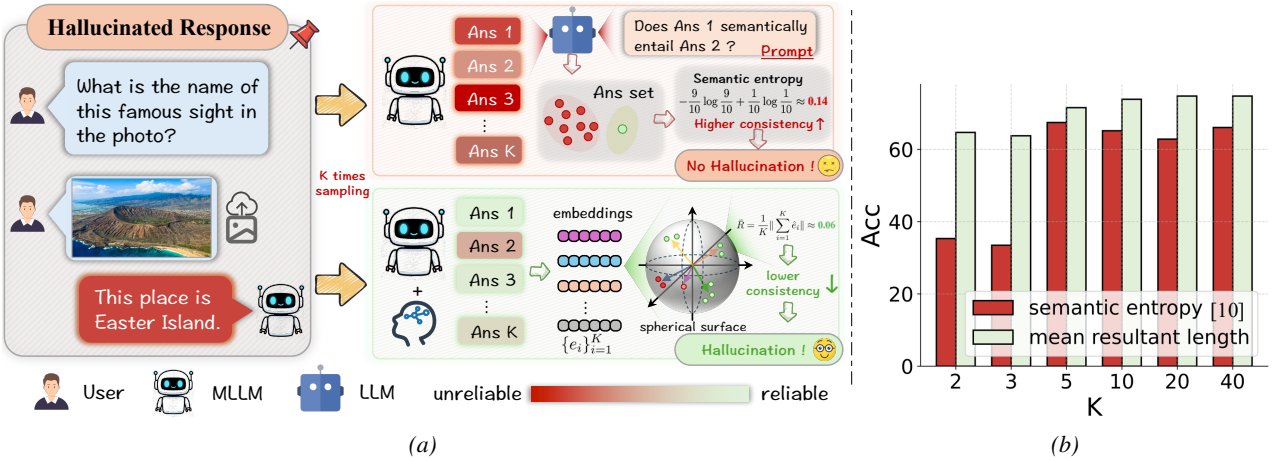


Figure 1. (a) Comparison of two ways to elicit model responses. If we directly query the model, it may generate many hallucinated answers with low uncertainty, leading to detection failure. In contrast, if we guide the model's responses, it is more likely to produce correct answers, making the uncertainty estimation more accurate. (b) Accuracy comparison of the two metrics under the same DOUBT framework. When the sample size K is small, the performance of semantic entropy (Farquhar et al., 2024) drops significantly and becomes highly unstable.

measured uncertainty and missed detections as shown in Figure 4. This stems from the modality gap where the visual module often lags behind the LLM's reasoning capabilities, leading to object-logic mismatches (Zhai et al., 2023). Second, existing methods commonly compute cluster-level semantic entropy to quantify the uncertainty scores by clustering multiple responses based on semantics and measuring cluster probabilities (Zhang et al., 2024; Kossen et al., 2024). However, semantic entropy is sensitive to clustering and unstable when the sample size K is small, as shown in Figure 1b.

To address the above limitations, we propose a model-agnostic approach that focuses on object-level hallucination detection, termed **D**ecoupled **O**bject-level **U**nderstanding and **B**ridging via vMF-based **T**rustworthiness (DOUBT). Our DOUBT proceeds in two stages. In the first stage, we address the limitations of the visual module by decoupling object recognition and relational understanding through a step-by-step reasoning paradigm, encouraging models to produce more diverse and accurate responses. More specifically, an MLLM is first prompted to i) list recognized objects in the input image, and ii) reason about the image conditioned on these objects using a Chain-of-Thought (CoT) prompt (Wei et al., 2022) to obtain new responses. In the end, we can sample the responses through both direct sampling and object-guided reasoning using the same input image to enhance the diversity of the response space, facilitating hallucination detection with only prompting strategies rather than modifying model architectures. In the second stage, to overcome the limitations of traditional semantic entropy in better assessing the uncertainty in the response set, we draw inspiration from the von Mises–Fisher (vMF) distribution (Fisher, 1953) and measure it using the mean resultant length, which is simple yet efficient to compute and can directly measure the similarity of input samples based on the geometric properties of the feature space. It proves to be more robust under the same conditions in Figure 1b. Once computed, we take the average uncertainty of the two types of responses and compare it with a predefined threshold to detect hallucinations.

Our main contributions are summarized as follows:

- This work reveals that the reasoning limitations of the visual module could cause the output results to remain incorrectly consistent, thus degrading the performance of hallucination detection. To address this, we decouple simple recognition from hard reasoning, and then guide the model to elicit richer and object-aware responses, thus improving hallucination detection.
- We propose a model-agnostic uncertainty metric based on the vMF distribution to alleviate the instability of traditional entropy measurement, which can be seamlessly integrated as a plug-and-play module into existing hallucination detection pipelines.
- We present extensive experiments across multiple MLLMs, scales, and four widely-used benchmarks (i.e., LLaVA-Bench, MM-Vet, MMMU, and ScienceQA), along with ablation studies demonstrating the superior accuracy and stability of our DOUBT.

2. Related Work

2.1. Hallucination Detection

Hallucination detection has attracted increasing attention from academia and industry as MLLMs are deployed in safety-critical scenarios (Sahoo et al., 2024; Moor et al.,

2023). Existing methods can be roughly grouped into three categories. i) Internal-signal methods analyze model-internal signals, such as activations, attentions, logits, or gradient-based uncertainty, to infer whether a response is hallucinated (Kadavath et al., 2022; Quevedo et al., 2025; Suresh et al., 2025). Although these methods can be effective, they require access to model internals, which is often inaccessible for proprietary or closed-source MLLMs (Dasgupta et al., 2025). ii) External-knowledge methods verify responses using external knowledge sources or fact-checking modules, ensuring consistency with objective information (Choi et al., 2023; Zhang et al., 2025). These methods are often reliable when relevant knowledge exists, but they incur additional computational overhead and depend on the completeness and quality of external resources (Sok et al., 2025). iii) Consistency-based methods detect hallucinations by comparing multiple model outputs, based on the intuition that correct answers tend to be self-consistent while hallucinations vary (Zhang et al., 2023; Kossen et al., 2024). Although broadly applicable and model-agnostic, these approaches can still fail when an MLLM repeatedly generates consistent but incorrect responses, particularly when the visual understanding module is weak (Chen et al., 2024a; Srey et al., 2025). In such cases, uncertainty appears seemingly low, and hallucinations remain undetected. To address this issue, this work proposes an object-level understanding and bridging approach to diversify and improve sampled responses, thereby ensuring that uncertainty estimates more accurately reflect true reliability.

2.2. Uncertainty Estimation

As a key tool for assessing model confidence and reliability, uncertainty estimation has emerged as an important topic within MLLM research (Kendall & Gal, 2017; He et al., 2023). It plays an important role in applications where robustness and trustworthiness are essential (Abdar et al., 2021; Gawlikowski et al., 2023). In the context of hallucination detection for MLLMs, uncertainty provides a practical proxy for whether model outputs should be trusted (Ovadia et al., 2019; Nguyen et al., 2025). Existing approaches for measuring uncertainty can be broadly divided into two main categories. Distribution-based methods rely on the predictive probability distribution of the model, with entropy and mutual information being the most common measures (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). These methods capture how concentrated or dispersed the output distribution is, but they often fail when models are overconfident in wrong predictions. Sampling-based methods, on the other hand, estimate uncertainty by generating multiple outputs and assessing their diversity (Kuleshov et al., 2018; Tian et al., 2025). Among these, semantic entropy (Kossen et al., 2024; Nikitin et al., 2024) has been widely used in this category, where the variance across responses is taken as a signal of uncertainty. Despite their effectiveness, entropy-based methods are highly sensitive to the number of samples and can be unstable in low-sample regimes. To address this, we propose a vMF-inspired trustworthiness metric that provides a smoother and more sample-efficient approximation of uncertainty, especially in low-sample regimes.

3. Method

3.1. Problem Statement

This work studies hallucination detection in a black-box setting, where no model internals or external knowledge bases are accessible. For the i-th multimodal input $p_i = \{I_i, t_i\}$, where I_i is the image and t_i is the task instruction (e.g., question or caption), an MLLM produces a response $A = \mathrm{MLLM}(p)$. Following prior work (Farquhar et al., 2024), we first obtain a deterministic reference answer $A_i = \mathrm{MLLM}(p_i)$ using a low temperature. To assess its reliability, we then query the same input K times: $s_{\text{ori}} = \{a_1, a_2, \ldots, a_K\}$, which forms a response set used to estimate uncertainty for hallucination detection. However, if the model exhibits limited visual understanding, it may repeatedly generate consistent but incorrect answers, leading to falsely low estimated uncertainty.

To mitigate this, we introduce an Object-level Understanding and Bridging (OUB) mechanism that prompts the model to explicitly decouple object recognition (i.e., understanding) from relational reasoning (i.e., bridging) during inference. This process yields an additional answer set s_{bri}, which not only enhances the model's understanding capability but also increases response diversity, providing a contrastive reference for evaluation (Section 3.2). After obtaining s_{ori} and s_{bri}, we estimate the trustworthiness score of each response set using a vMF-based trustworthiness metric and compute their average (Section 3.3). The average score is then compared with a predefined threshold θ to determine whether the model's response is reliable, thereby enabling effective hallucination detection.

3.2. Object-level Understanding and Bridging

To improve the reasoning ability of the model, we design a two-stage prompting scheme that produces a set of bridging responses s_{bri}.

Stage 1: Object-level Understanding We begin with an object recognition prompt p_{obj} that instructs the model to identify the salient visual entities in the image. Objects serve as the fundamental building blocks of visual semantics, which provide the primary cues for understanding what the scene contains and guiding reasoning about the image.

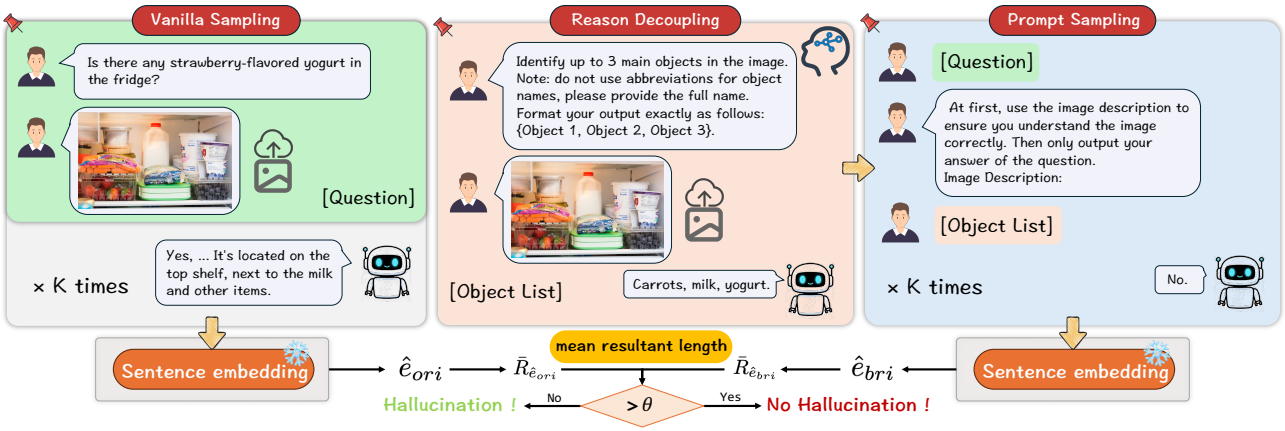


Figure 2. Overview of our DOUBT. Given an input, we obtain two response sets: direct responses and object-guided responses via decoupled prompting. Each set is embedded and scored using our vMF-based trustworthiness metric. The average score is then compared to a threshold to determine hallucination.

In other words, the correlations among these prominent objects generally determine the overall meaning of the image, guiding subsequent reasoning about actions, attributes, and higher-level contextual relationships. Establishing an accurate inventory of objects through p_{obj} is therefore critical for grounded reasoning. We empirically limit p_{obj} to 'up to 3 main objects' to balance attention focus with coverage. The prompt p_{obj} is shown as follows:

> *"Identify up to 3 main objects in the image. Note: do not use abbreviations for object names, please provide the full name. Format your output exactly as follows: {Object 1, Object 2, Object 3}."*

Mathematically, the object recognition result r_i is then obtained as:

$$r_i = \text{MLLM}(\{I_i, p_{\text{obj}}\}). \tag{1}$$

Stage 2: Object-level Bridging Next, we design a bridging prompt p_{bri} that guides the model to reason about the image conditioned on the recognized objects, and then generate an answer. This step encourages the model to verify and refine its understanding based on explicit object information, thereby reducing the likelihood of generating spurious or hallucinated responses. The full bridging prompt p_{bri} is given as:

> *"At first, use the image description to ensure you understand the image correctly. Then only output your answer of the question. \nImage Description: {Object 1, Object 2, Object 3}."*

With p_{bri}, we prompt the model to generate K object-bridged responses to form the bridging answer set s_{bri} as below: $s_{\text{bri}} = \{b_1, b_2, \ldots, b_K\}$, where $b_i = \text{MLLM}(\{I_i, t_i || p_{\text{bri}} || r_i\})$.

Through this two-step prompting, the model not only follows the user's instruction but also anchors reasoning on explicit object-level semantics, thus expanding the diversity and reliability of the answer space, facilitating trustworthy hallucination detection. We note that DOUBT relies on the model's object recognition capability, and severe object misidentification may propagate to later stages. However, since both stages are driven by the same MLLM, such failures typically manifest as increased response inconsistency, which our trustworthiness metric is designed to capture. Regarding complexity, while the two-step prompting linearly increases inference cost, it eliminates the substantial training overhead of white-box methods, making it a cost-effective trade-off for high-stakes black-box deployment.

3.3. vMF-Based Trustworthiness

A key indicator of hallucinations is the semantic divergence among multiple responses: dispersed answers suggest ambiguity in the model's understanding and thus a higher risk of hallucination. However, traditional entropy-based measures heavily rely on the accuracy of semantic clustering, which is unstable under small sample sizes. To overcome this, we present a geometric trustworthiness metric inspired by the vMF distribution.

The probability density function of the vMF distribution is:

$$p(x|\mu, \kappa) = C_d(\kappa) \exp(\kappa \mu^T x), \tag{2}$$

where $x \in \mathbb{R}^d$ is a unit vector, $\mu \in \mathbb{R}^d$ is the mean direction with $\|\mu\| = 1$, $\kappa \geq 0$ is the concentration parameter, and $C_d(\kappa)$ is a normalization constant in d dimensions.

On the unit spherical surface, response embeddings following a vMF distribution exhibit concentration governed by κ. Since κ is unbounded and difficult to threshold directly, we instead use the sample mean resultant length

$\bar{R} = \frac{1}{K}\|\sum_{i=1}^{K} \hat{x}_i\|$, which is an unbiased estimate related to κ according to the directional statistics theory (Mardia & Jupp, 2009):

$$\bar{R} = A_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}, \tag{3}$$

where $A_d(\kappa)$ denotes the ratio of modified Bessel functions that characterizes the mean resultant length in d-dimensional space; $I_\nu(\cdot)$ is the modified Bessel function of the first kind. Like κ, $\bar{R}$ can be directly used to measure the concentration of the response distribution. Specifically, a larger $\bar{R}$ is equivalent to a larger κ in the vMF distribution, indicating higher response consistency and lower hallucination risk.

To compute $\bar{R}$, we encode each textual response using the `nli-roberta-large` model, yielding embedding sets: $E_{\text{ori}} = \{e_{a_1}, e_{a_2}, \ldots, e_{a_K}\}$ and $E_{\text{bri}} = \{e_{b_1}, e_{b_2}, \ldots, e_{b_K}\}$, where $e_{a_i} = \text{nrl}(a_i)$ and $e_{b_i} = \text{nrl}(b_i)$. Since the vMF distribution is defined over directional data, i.e., points lying on the surface of a unit hypersphere, we normalize E_{ori} and E_{bri} onto a unit hypersphere as follows: $\hat{E}_{\text{ori}} = \{\hat{e}_{a_i}\}_{i=1}^{K} = \{\frac{e_{a_i}}{\|e_{a_i}\|}\}_{i=1}^{K}$ and $\hat{E}_{\text{bri}} = \{\hat{e}_{b_i}\}_{i=1}^{K} = \{\frac{e_{b_i}}{\|e_{b_i}\|}\}_{i=1}^{K}$. We normalize embeddings onto a unit hypersphere because, in high-dimensional semantic spaces (e.g., RoBERTa (Liu et al., 2019)), semantic similarity is primarily encoded in direction (angular distance) rather than magnitude, which often correlates with token frequency noise. Thus, vMF offers a more theoretically grounded measure for semantic consistency than Euclidean-based metrics.

After normalization, the mean resultant lengths for the two response sets (i.e., $\hat{E}_{\text{ori}}$ and $\hat{E}_{\text{bri}}$) are respectively computed as: $\bar{R}_{\hat{E}_{\text{ori}}} = \frac{1}{K}\|\sum_{i=1}^{K} \hat{e}_{a_i}\|$ and $\bar{R}_{\hat{E}_{\text{bri}}} = \frac{1}{K}\|\sum_{i=1}^{K} \hat{e}_{b_i}\|$. Finally, we average them to obtain the overall trustworthiness score:

$$\bar{R}_{\text{avg}} = \frac{1}{2}(\bar{R}_{\hat{E}_{\text{ori}}} + \bar{R}_{\hat{E}_{\text{bri}}}). \tag{4}$$

A higher $\bar{R}_{\text{avg}}$ indicates greater internal consistency among responses and therefore higher confidence in the model's reliability.

3.4. Hallucination Detection Criterion

We determine the presence of hallucination by comparing the average trustworthiness score $\bar{R}_{\text{avg}}$ against a given threshold θ. Formally, the hallucination flag is given by:

$$flag = \begin{cases} 1, & \text{if } \bar{R}_{\text{avg}} > \theta, \\ 0, & \text{if } \bar{R}_{\text{avg}} \leq \theta, \end{cases} \tag{5}$$

where $flag = 1$ denotes that the model's responses exhibit sufficiently high internal trustworthiness and are therefore judged as non-hallucinatory, while $flag = 0$ denotes low trustworthiness and is treated as a hallucination case. This decision rule aligns with the intuition that reliable answers should be directionally coherent across multiple reasoning paths, whereas hallucinations typically result in semantic divergence.

Finally, we evaluate detection accuracy by comparing the reference answer A_i against its ground-truth answer gt_i. More specifically, we compare the predicted hallucination flag with the ground-truth correctness of the reference answer A_i:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\mathbb{I}[A_i = gt_i] = flag\right], \tag{6}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function and N is the total number of evaluation samples. Generally, a detection is considered correct if the predicted hallucination flag aligns with the actual correctness of the answer, and the overall accuracy measures the proportion of correctly detected samples.

4. Experiment

4.1. Experimental Setup

Datasets. We evaluate our method on widely used benchmarks covering both free-form and multiple-choice settings. For free-form evaluation, we use LLaVA-Bench (Liu et al., 2023) and MM-Vet (Yu et al., 2024), which assess high-level visual reasoning and integrated vision-language understanding. For multiple-choice evaluation, we adopt MMMU (Yue et al., 2024), a challenging college-level benchmark, and ScienceQA (Lu et al., 2022), which evaluates factual knowledge and systematic reasoning.

Models. We evaluate ten representative MLLMs from four architectural families: Qwen2VL, InternVL2, LLaVA-1.5, and LLaVA-NeXT. Specifically, the Qwen2VL series (i.e., 2B, 7B, and 72B) (Wang et al., 2024a) adopts an end-to-end vision-language architecture; the InternVL2 series (i.e., 1B, 8B, and 26B) (Chen et al., 2024b) emphasizes high-resolution visual encoding and cross-modal alignment; the LLaVA-1.5 series (i.e., 7B and 13B) (Liu et al., 2023) employs a lightweight CLIP-to-LLM projection (Radford et al., 2021); and the LLaVA-NeXT series (i.e., 7B and 13B) (Liu et al., 2024b) further improves resolution and training efficiency.

Implementation Details. Following prior works (Chen et al., 2024a; Farquhar et al., 2024), the parameter configuration for all models in the experiments is set to *temperature* = 0.1 when getting the reference answer, and *temperature* = 0.5, *top-k* = 10, *top-p* = 0.99 when generating responses for uncertainty evaluation. The sampling time K is set to 10. The threshold θ we use to measure the confidence of responses is 0.48 based on the parameter analysis in Figure 3c.

Baselines. We compare our method with six represen-

Dataset	Method	Q2B	Q7B	Q72B	I1B	I8B	I26B	L7B	L13B	LN7B	LN13B	Avg
LLaVABench	GAVIE (2024)	25.00	26.67	40.00	30.00	31.67	31.67	15.00	20.00	45.00	35.00	30.00
	Semantic Entropy (2024)	61.67	55.00	61.67	65.00	60.00	53.33	70.00	**70.00**	61.67	**65.00**	62.33
	KLE (2024)	28.33	48.33	58.33	40.00	46.67	50.00	23.33	45.00	43.33	33.33	41.67
	EigenScore (2024)	63.33	**58.33**	**63.33**	63.33	55.00	**61.67**	68.33	**70.00**	**70.00**	**65.00**	63.83
	VL-Uncertainty (2025)	55.67	53.33	53.33	60.00	60.00	51.67	73.33	63.33	61.67	61.67	59.40
	SNNE (2025)	45.00	56.67	55.00	38.33	53.33	56.67	31.67	41.67	41.67	46.67	46.67
	Ours	**68.33**	**58.33**	53.33	**73.33**	**61.67**	55.00	**80.00**	66.67	63.33	**65.00**	**64.50**
MMVet	GAVIE (2024)	29.36	43.58	51.38	30.73	30.73	22.48	23.39	24.77	37.61	43.58	33.76
	Semantic Entropy (2024)	60.55	57.80	62.84	72.94	55.05	58.72	72.48	79.36	61.01	72.48	65.32
	KLE (2024)	45.41	51.83	56.88	42.20	46.79	51.38	41.74	41.28	41.74	42.66	46.19
	EigenScore (2024)	**73.85**	70.64	**72.48**	77.98	67.43	**76.61**	73.85	78.44	73.85	77.52	74.27
	VL-Uncertainty (2025)	64.22	67.43	66.97	65.60	62.39	64.67	**79.35**	**80.28**	66.06	69.72	68.67
	SNNE (2025)	48.62	57.80	63.30	40.37	49.54	54.13	39.45	42.20	43.12	48.62	48.72
	Ours	**73.85**	**72.94**	71.10	**78.44**	**72.02**	75.23	76.61	77.98	**76.15**	**77.98**	**75.23**
MMMU	GAVIE (2024)	37.82	48.36	57.09	40.61	48.12	33.21	37.58	44.61	43.64	45.82	43.69
	Semantic Entropy (2024)	53.82	54.91	60.36	53.82	54.91	52.48	52.61	50.18	52.61	50.18	53.59
	KLE (2024)	43.88	53.33	62.91	46.42	49.58	59.52	51.03	45.33	47.39	51.52	51.09
	EigenScore (2024)	51.43	64.48	67.64	51.15	**59.39**	56.24	**59.03**	54.42	55.39	49.21	56.84
	VL-Uncertainty (2025)	57.33	58.55	65.94	55.15	57.33	57.21	56.36	55.15	57.58	**56.24**	57.68
	SNNE (2025)	43.52	53.33	62.79	48.24	47.27	59.27	40.73	49.58	45.94	55.52	50.62
	Ours	**60.84**	**64.61**	**68.72**	**57.58**	**59.39**	**59.64**	58.55	**56.48**	**60.85**	55.52	**60.22**
ScienceQA	GAVIE (2024)	61.82	77.09	85.23	53.94	86.71	89.19	58.50	66.39	62.27	65.20	70.63
	Semantic Entropy (2024)	54.04	77.94	87.06	64.45	**90.08**	91.32	61.77	68.02	67.67	65.34	72.77
	KLE (2024)	62.22	76.45	86.91	67.63	89.64	90.43	60.73	67.97	65.94	66.23	73.42
	EigenScore (2024)	62.57	72.73	84.63	62.22	57.51	87.31	64.15	65.64	70.80	64.95	69.25
	VL-Uncertainty (2025)	**66.83**	**80.71**	88.60	64.50	89.54	**91.57**	65.79	68.57	68.67	**67.67**	75.25
	SNNE (2025)	64.35	78.63	84.68	66.04	79.18	83.14	65.54	68.02	70.70	67.50	72.78
	Ours	65.00	79.87	**88.80**	**68.32**	**90.08**	91.37	**66.48**	**72.43**	**70.95**	67.43	**76.07**

Table 1. Comparison between our method and other state-of-the-art baselines. Q, I, L, and LN denote Qwen2VL, InternVL2, LLaVA-1.5, and LLaVA-NeXT, respectively. The reported results are detection accuracies in percentage. **Bold** numbers indicate the best performance, while underlined numbers represent the second best.

tative baselines, i.e., GAVIE (Liu et al., 2024a) (external LLM-based factuality check), Semantic Entropy (Farquhar et al., 2024) (uncertainty over semantic clusters), KLE (Nikitin et al., 2024) (kernel-based continuous entropy), EigenScore (Chen et al., 2024a) (embedding diversity via covariance eigenvalues), VL-Uncertainty (Zhang et al., 2024) (probability-based uncertainty estimation) and SNNE (Nguyen et al., 2025) (entropy incorporating pairwise semantic similarity). These baselines cover a diverse range of mainstream hallucination detection paradigms, including entropy-based, sampling-based, and metric-based approaches, ensuring a comprehensive and fair evaluation.

4.2. Comparison with State-of-the-Art Baselines

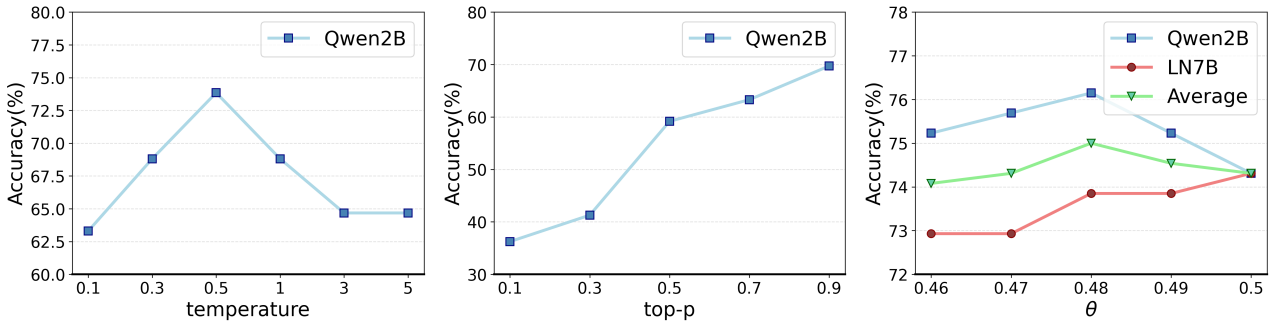
Table 1 reports the experimental results across four multimodal benchmarks. From the results, we can see that our DOUBT consistently achieves superior or competitive performance compared to all state-of-the-art baselines.

Free-form Datasets. On the LLaVABench dataset, our method achieves an average accuracy of 64.50%, ranking second overall while slightly outperforming EigenScore's 63.83%. Notably, our DOUBT achieves the highest improvement of 8.33% when using the InternVL2-1B model, showing particularly strong adaptation to smaller-scale models. Moreover, gains are more pronounced on L7B than on L13B. We attribute this to larger models possessing stronger inherent reasoning, thus benefiting less from explicit OUB guidance compared to smaller models. Results on the MM-Vet dataset show our method attaining the highest average accuracy of 75.23%, significantly surpassing EigenScore's 74.27%. The method maintains consistent leadership across the InternVL2 model series, achieving 78.44%, 72.02%, and 75.23% on the I1B, I8B, and I26B models, respectively, demonstrating effective scalability across different model sizes.

Multiple-choice Datasets. Our method also exhibits clear advantages on multiple-choice benchmarks. Specifically,

Dataset	Configuration	Q2B	Q7B	Q72B	I1B	I8B	I26B	L7B	L13B	LN7B	LN13B	Avg
MMMU	Full version	60.84	**64.61**	**68.72**	57.58	59.39	59.64	58.55	56.48	**60.85**	55.52	**60.22**
	- w/o OUB	**61.94**	62.18	66.30	55.27	59.64	**59.88**	59.27	57.94	59.88	55.39	59.77
	- w/o vMF	60.96	59.35	67.15	**59.61**	**60.73**	55.39	**59.76**	**59.96**	58.84	**58.38**	60.01
	- w/o OUB & vMF	56.73	58.91	59.88	57.09	54.55	55.15	56.97	56.36	57.58	53.82	56.70
ScienceQA	Full version	65.00	**79.87**	**88.80**	**68.32**	**90.08**	91.37	66.48	**72.43**	70.95	67.43	**76.07**
	- w/o OUB	65.99	79.03	88.00	68.12	89.99	**91.52**	**66.98**	70.90	**71.15**	67.58	75.93
	- w/o vMF	**66.21**	65.47	61.63	63.99	69.33	81.11	63.91	64.71	66.04	**68.57**	67.10
	- w/o OUB & vMF	57.26	63.66	68.82	57.11	67.38	68.66	58.55	59.74	60.29	58.80	62.03

Table 2. Ablation study of our method on two datasets by removing two key components. We compare the full method with two variants: without object-level understanding and bridging (- w/o OUB) and without vMF-based trustworthiness (- w/o vMF). The reported numbers are detection accuracies in percentage. **Bold** numbers indicate the best performance.



(a) *(b)* *(c)*

Figure 3. (a) Performance sensitivity to temperature. (b) Performance sensitivity to top-p. (c) Performance sensitivity to threshold.

on the MMMU dataset, DOUBT achieves a notable average accuracy of 60.22%, exceeding VL-Uncertainty's 57.68% by a considerable margin of 2.54 percentage points. The method reaches its excellent performance of 68.72% using the Qwen2VL-72B model, highlighting its advantage with larger-scale models. On the ScienceQA dataset, DOUBT achieves the highest average accuracy of 76.07%, outperforming VL-Uncertainty's 75.25%. The approach achieves consistently better performance than baseline methods across most models, demonstrating strong capability in scientific question answering tasks.

Overall, our DOUBT achieves the best average performance on all benchmarks. It maintains consistent performance across various model scales, ranging from 1B to 72B parameters. Other baseline methods fail to achieve consistently strong performance across all datasets. While some methods (e.g., EigenScore) may perform well on individual datasets or models, their performance varies across models and scales. In contrast, DOUBT not only surpasses strong baselines with large margins on datasets like LLaVABench and MM-Vet, but also maintains stable competitiveness on more difficult benchmarks such as MMMU and ScienceQA, demonstrating its superior generalization and robustness. This is attributed to our use of the OUB strategy and vMF-based trustworthiness, which makes the detection results more accurate and stable compared to other methods that rely solely on multiple sampling and entropy-based measures.

4.3. Ablation Study

To verify the contribution of each component, we conduct detailed ablation experiments on the larger MMMU and ScienceQA datasets to reduce the randomness caused by limited data size. Table 2 reports the ablation results, where two key components are removed: OUB and vMF-based Trustworthiness. From the results, we observe that removing either component may improve performance in certain test scenarios, but overall performance is negatively affected. For example, after removing OUB, the Qwen2VL-2B model shows improved performance on the MMMU dataset, and removing the vMF-based Trustworthiness also leads to better results on the ScienceQA dataset. However, the full DOUBT method consistently achieves the highest average scores (i.e., 60.22% on MMMU and 76.07% on ScienceQA), surpassing all ablated variants. These results demonstrate that both components are essential and complementary: OUB enriches response diversity and semantic grounding, while vMF-based Trustworthiness provides a stable and geometry-aware measure of reliability.

Question: What is the name of the man in the photo?
GTAns: The man in the photo is Elon Musk.
MLLMAns: I don't know who the man in the photo is.
Original Answer Set:
- I don't know who this person is.
- I don't know who the man in the photo is.
- I don't know who the man in the photo is.
- I'm unable to identify or name individuals in images.
- I don't know who the man in the photo is.
- I don't know who this person is.
- I don't know who the man in the photo is.
- I don't know who the man in the image is.
- I don't know who the man in the photo is.
- The image is an edited artwork and does not represent an actual person. It features a character that is a combination of a human and a chimpanzee, holding a dog. This is not a photograph of an actual person, so I cannot identify the name of the man in the photo.
Object List: {Man, Dog, Clouds}
OLB Answer Set:
- Elon Musk
- The image shows Elon Musk, the CEO of SpaceX and Tesla, holding a dog in a pose that resembles a space-suited astronaut. The dog is depicted as a Shiba Inu, which is a popular dog breed. The background appears to be a cloudy sky, adding to the surreal and imaginative nature of the scene. Elon Musk is known for his innovative work in space exploration and technology, which might be the inspiration behind this whimsical image.
- The man in the photo appears to be Elon Musk.
- Elon Musk
- The image description does not provide any specific information about the name of the man in the photo.
- I cannot identify people in images.
- Elon Musk
- The image shows a man with a dog, set against a backdrop of clouds. However, I cannot identify or name individuals in images.
- Elon Musk
- Elon Musk

Semantic Entropy: -0.0 ✗
EigenScore: -1.768 ✗
DOUBT: 0.314 ✓

Question: Is there any strawberry-flavored yogurt in the fridge?

Semantic Entropy: -0.0 ✗
EigenScore: -2.519 ✗
DOUBT: 0.300 ✓

GTAns: There is no strawberry-flavored yogurt in the fridge. There is a large bottle of Fage non-fat yogurt, a smaller cup of Fage blueberry yogurt, and another smaller cup with an unknown brand and flavor.
MLLMAns: Yes, there is a container of strawberry-flavored yogurt in the fridge.
Original Answer Set:
- Yes, there is a container of strawberry-flavored yogurt in the fridge.
- Yes, there is a container of strawberry-flavored yogurt in the refrigerator.
- Yes, there is a container of strawberry-flavored yogurt in the fridge.
- Yes, there is a container of strawberry-flavored yogurt in the fridge.
- Yes, there is a container of strawberry-flavored yogurt in the refrigerator.
- Yes, there is a container of strawberry-flavored yogurt in the fridge.
- Yes, there is a container of strawberry-flavored yogurt in the refrigerator.
- Yes, there is a container of strawberry-flavored yogurt in the fridge.
- Yes, there is a container of strawberry-flavored yogurt in the fridge.
- Yes, there is a container of strawberry-flavored yogurt in the fridge.
Object List: {Carrots, milk, yogurt}
OLB Answer Set:
- No - Yes - Yes - No - No - No - Yes - Yes - No - No

Figure 4. Examples of hallucination detection cases. Compared to our method, the other approaches tend to produce falsely consistent results and achieve lower hallucination metrics.

4.4. Parameter Study

We conduct a sensitivity analysis on answer-generation hyperparameters, including temperature, top-p, and threshold, using Qwen2VL-2B on MM-Vet, with LLaVA-NeXT-7B additionally evaluated for threshold analysis. As shown in Figure 3, the model exhibits distinct behaviors under different hyperparameter settings. Temperature exhibits a non-monotonic effect, with performance first improving and then saturating, while top-p shows a clear positive correlation with accuracy. In contrast, performance is relatively stable across different threshold values for both models, with the best overall performance achieved at a threshold of 0.48. Overall, top-p has the most significant impact on accuracy, whereas temperature and threshold introduce only moderate variations.

4.5. Case Study

We further illustrate the interpretability and reliability of DOUBT with qualitative examples in Figure 4. In both cases, the model initially produces incorrect answers, which can mislead existing detection methods. With DOUBT, the model first performs object-level understanding to identify key entities, and then applies object-level bridging to re-reason the question grounded on these entities. As a result, the generated responses become more consistent with the visual evidence, leading to more reliable trustworthiness estimates. These examples demonstrate how OUB diversifies the response space and refines trustworthiness estimation.

5. Conclusion

In this paper, we investigate hallucination detection in MLLMs, a critical step toward ensuring trustworthy and safe deployment. We propose DOUBT, a black-box reliability assessment framework that integrates step-by-step reasoning prompting and geometric trustworthiness modeling. Specifically, the proposed Object-level Understanding and Bridging (OUB) mechanism decouples visual reasoning from relational reasoning, enabling more diverse and semantically grounded answers. Meanwhile, the vMF-based Trustworthiness metric leverages the mean resultant vector length of the response embeddings to measure trustworthiness in a stable and geometry-aware manner, offering a superior alternative to traditional entropy-based approaches. Extensive experiments across four benchmarks and ten MLLMs demonstrate DOUBT's superior performance and consistent generalization across model scales from 1B to 72B parameters. Unlike previous methods that rely solely on sampling or entropy measures, our DOUBT provides a theoretically grounded and practically robust approach to hallucination detection. We hope our findings inspire future research toward interpretable, uncertainty-aware multimodal reasoning and contribute to the development of more reliable MLLMs.

Broader Impact

This work aims to improve the reliability of MLLMs by detecting hallucinated outputs. By providing trustworthiness estimates rather than directly modifying model behavior, our approach can support safer deployment of MLLMs in downstream applications. While the method may be used in high-stakes settings, it does not automate decision-making and is intended to complement, rather than replace, human judgment.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ahadian, P. and Guan, Q. A survey on hallucination in large language and foundation models. *Preprints*, April 2025. doi: 10.20944/preprints202504.1236.v1. URL https://doi.org/10.20944/preprints202504.1236.v1.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., and Shou, M. Z. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., and Ye, J. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=Zj12nzlQbz.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Choi, S., Fang, T., Wang, Z., and Song, Y. KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14035–14053, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.867. URL https://aclanthology.org/2023.emnlp-main.867/.

Dasgupta, S., Nath, S., Basu, A., Shamsolmoali, P., and Das, S. Hallushift: Measuring distribution shifts towards hallucination detection in llms. *arXiv preprint arXiv:2504.09482*, 2025.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Fisher, R. A. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., and He, R. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=DgH9YCsqWm.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

He, W., Jiang, Z., Xiao, T., Xu, Z., and Li, Y. A survey on uncertainty quantification methods for deep learning. *arXiv preprint arXiv:2302.13425*, 2023.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Huang, W., Liu, H., Guo, M., and Gong, N. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9614–9631, 2024.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Li, Q., Geng, J., Lyu, C., Zhu, D., Panov, M., and Karray, F. Reference-free hallucination detection for large vision-language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4542–4551, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.262. URL https://aclanthology.org/2024.findings-emnlp.262/.

Li, Q., Geng, J., Chen, Z., Zhu, D., Wang, Y., Ma, C., Lyu, C., and Karray, F. HD-NDEs: Neural differential equations for hallucination detection in LLMs. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6173–6186, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.309. URL https://aclanthology.org/2025.acl-long.309/.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=J44HfH4JCg.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024b.

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., and Peng, W. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253, 2024c. URL https://doi.org/10.48550/arXiv.2402.00253.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Manakul, P., Liusie, A., and Gales, M. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Mardia, K. V. and Jupp, P. E. *Directional statistics*. John Wiley & Sons, 2009.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., and Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Nguyen, D., Payani, A., and Mirzasoleiman, B. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. *In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Nikitin, A., Kossen, J., Gal, Y., and Marttinen, P. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Quevedo, E., Salazar, J. Y., Koerner, R., Rivas, P., and Cerny, T. Detecting hallucinations in large language model generation: A token probability approach. In Arabnia, H. R., Deligiannidis, L., Amirian, S., Shenavarmasouleh, F., Ghareh Mohammadi, F., and de la Fuente, D. (eds.), *Artificial Intelligence and Applications*, pp. 154–173, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-86623-4.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., and Chadha, A. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11709–11724, 2024.

Sok, C., Luz, D., and Haddam, Y. Metarag: Metamorphic testing for hallucination detection in rag systems. *arXiv preprint arXiv:2509.09360*, 2025.

Srey, P., Wu, X., and Luu, A. T. Unsupervised hallucination detection by inspecting reasoning processes. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22128–22140, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1124. URL https://aclanthology.org/2025.emnlp-main.1124/.

Suresh, M., Aljundi, R., Nkisi-Orji, I., and Wiratunga, N. Cross-layer attention probing for fine-grained hallucination detection. *arXiv preprint arXiv:2509.09700*, 2025.

Tian, Y., Jin, P., Yuan, M., Li, N., Zeng, B., and Li, Q. RODS: Robust optimization inspired diffusion sampling for detecting and reducing hallucination in generative models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=fhuqIxoPcr.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024a. URL https://doi.org/10.48550/arXiv.2409.12191.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., XiXuan, S., et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024b.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Zhai, B., Yang, S., Xu, C., Shen, S., Keutzer, K., Li, C., and Li, M. Halle-control: controlling object hallucination in large multimodal models. *arXiv preprint arXiv:2310.01779*, 2023.

Zhang, J., Li, Z., Das, K., Malin, B., and Kumar, S. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15445–15458, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.

1032. URL https://aclanthology.org/2023.findings-emnlp.1032/.

Zhang, J., Xu, C., Gai, Y., Lecue, F., Yang, S., Song, D., and Li, B. Knowhalu: Hallucination detection via multi-form knowledge based factual checking. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. URL https://openreview.net/forum?id=RFwyhpcYZK.

Zhang, R., Zhang, H., and Zheng, Z. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

A. Hallucination Mitigation through OUB

Dataset	Configuration	Q2B	Q7B	Q72B	I1B	I8B	I26B	L7B	L13B	LN7B	LN13B	Avg
MME	Initial	82.22	**88.33**	**91.53**	54.17	71.02	**79.91**	76.71	**74.77**	58.00	**73.63**	75.03
	With OUB	**82.31**	**88.33**	**91.53**	**54.21**	**72.24**	79.87	**76.79**	74.43	**62.55**	73.17	**75.54**
POPE (random)	Initial	89.77	**88.63**	**89.70**	67.47	80.63	72.57	75.60	87.63	56.40	87.13	79.55
	With OUB	**89.90**	88.60	89.60	**67.90**	**81.40**	**73.13**	**77.37**	**88.67**	**59.33**	**88.27**	**80.42**
POPE (adversarial)	Initial	**87.37**	**86.20**	**86.37**	**65.13**	77.47	69.07	71.87	79.80	49.27	**83.40**	75.60
	With OUB	87.30	86.13	86.33	64.33	**78.27**	**70.10**	**72.57**	**80.3**	**51.17**	82.17	**75.87**
POPE (popular)	Initial	88.37	**87.50**	**88.33**	66.53	79.07	70.87	74.73	84.70	49.30	**86.00**	77.54
	With OUB	**88.53**	87.47	88.20	**66.60**	**79.50**	**71.50**	**75.57**	**85.23**	**51.33**	85.73	**77.97**

Table 3. Performance of hallucination mitigation achieved through OUB. The results are reported as answer accuracy in percentage. **Bold** numbers indicate the best performance.

To examine the effectiveness of OUB in correcting hallucinated responses, we conduct experiments on the POPE (Li et al., 2023b) and MME (Fu et al., 2025) datasets. We first prompt the model to generate multiple responses and then use the vMF-based trustworthiness metric to decide whether to apply OUB for hallucination mitigation. When necessary, OUB guides the model to generate a caption for the image, identify the objects mentioned in the caption, and finally produce a refined answer based on both the caption and the object list. Table 3 presents the answer accuracy of different models before and after applying OUB. The results show that with OUB, models achieve higher accuracy in most cases, indicating that OUB effectively mitigates hallucination. Therefore, we incorporate OUB into DOUBT to expand the diversity of the model's answer space, making uncertainty detection more reliable and consequently improving hallucination detection performance.

B. Performance under Different vMF Parameters

Dataset	Configuration	Q2B	Q7B	Q72B	I1B	I8B	I26B	L7B	L13B	LN7B	LN13B	Avg
MMVet	Ours	**73.85**	**72.94**	**71.10**	**78.44**	72.02	75.23	**76.61**	**77.98**	**76.15**	**77.98**	**75.23**
	Ours (κ)	72.48	72.48	70.18	77.98	**72.48**	**77.06**	**76.61**	77.52	**76.15**	77.52	75.05
MMMU	Ours	60.84	**64.61**	**68.72**	**57.58**	**59.39**	59.64	**58.55**	**56.48**	60.85	**55.52**	**60.22**
	Ours (κ)	**60.97**	63.88	**68.72**	55.88	58.91	**59.88**	57.21	55.76	**60.97**	55.03	59.72

Table 4. Ablation study of our method on MM-Vet and MMMU datasets under different vMF parameter settings. Results are reported for both $\bar{R}$ and κ, with detection accuracies given in percentage. **Bold** numbers indicate the best performance.

Table 4 presents the results obtained using different parameters in vMF, where after normalizing κ we set the threshold to 0.38. It can be observed that the overall performance gap between using $\bar{R}$ and κ is not substantial, especially in terms of the average scores across datasets. This observation is consistent with the mathematical relationship discussed earlier between the two parameters. However, in terms of overall performance, using κ is still somewhat worse than using $\bar{R}$, which further confirms our earlier point that the wide range of κ makes it challenging to determine an effective threshold.

C. Effect of Introducing a More Capable External Model

For a further investigation of the effect of OUB, we replace the model's own object list with that provided by the Qwen-VL-Plus model (Bai et al., 2023). Table 5 presents a comparison of hallucination detection performance before and after applying OUB with a more capable external model. It can be observed that in most cases, the detection performance improves, indicating that when the model performs OUB using more accurate information, it is better able to produce correct answers and the results align with our expectations.

Dataset	Configuration	Q2B	I8B	L13B	LN7B	Avg
LLaVABench	Ours	68.33	61.67	66.67	63.33	65.00
	With QVP	**71.67**	**63.33**	**70.00**	**65.00**	**67.50**
MMVet	Ours	**73.85**	72.02	77.98	**76.15**	75.00
	With QVP	72.02	**74.77**	**81.19**	**76.15**	**76.03**
ScienceQA	Ours	65.00	90.08	**72.43**	70.95	74.62
	With QVP	**66.39**	**90.38**	71.19	**72.68**	**75.16**

Table 5. Ablation study on the effect of introducing a more capable model. QVP denotes Qwen-VL-Plus. Results are reported as detection accuracies in percentage. **Bold** numbers indicate the best performance.

D. Impact of Different Uncertainty Estimation Metrics

Dataset	Metric	Q2B	Q7B	Q72B	I1B	I8B	I26B	L7B	L13B	LN7B	LN13B	Avg
LLaVABench	SE	**68.33**	**58.33**	55.00	71.67	**63.33**	53.33	**80.00**	63.33	**65.00**	63.33	64.17
	ES	48.33	**58.33**	**60.00**	50.00	60.00	**61.67**	50.00	**66.67**	60.00	61.67	57.67
	vMF-T	**68.33**	**58.33**	53.33	**73.33**	61.67	55.00	**80.00**	**66.67**	63.33	**65.00**	**64.50**
MMVet	SE	68.46	70.85	59.17	71.25	71.23	69.50	69.91	71.67	73.85	69.57	69.55
	ES	69.72	**76.61**	**72.02**	75.69	**73.85**	**77.06**	76.15	**80.28**	73.31	77.06	75.18
	vMF-T	**73.85**	72.94	71.10	**78.44**	72.02	75.23	**76.61**	77.98	**76.15**	**77.98**	**75.23**
MMMU	SE	**60.96**	59.35	67.15	**59.61**	60.73	55.39	**59.76**	**59.96**	58.84	**58.38**	60.01
	ES	54.06	**66.55**	**68.85**	53.58	**61.09**	55.52	56.48	56.24	56.61	50.67	57.97
	vMF-T	60.84	64.61	68.72	57.58	59.39	**59.64**	58.55	56.48	**60.85**	55.52	**60.22**
ScienceQA	SE	**66.21**	65.47	61.63	63.99	69.33	81.11	63.91	64.71	66.04	**68.57**	67.10
	ES	61.87	72.24	85.57	56.37	59.74	85.72	63.11	68.77	70.90	63.41	68.77
	vMF-T	65.00	**79.87**	**88.80**	**68.32**	**90.08**	**91.37**	**66.48**	**72.43**	**70.95**	67.43	**76.07**

Table 6. Comparison of different uncertainty estimation metrics. SE, ES, and vMF-T denote Semantic Entropy, EigenScore and vMF-based Trustworthiness. Results are reported as detection accuracies in percentage. **Bold** numbers indicate the best performance.

When calculating uncertainty, the specific metric used has a significant impact on the final results. To compare the performance of different metrics, we apply each metric to DOUBT and examine the outcomes. Table 6 presents the results of using different uncertainty estimation metrics in DOUBT. It can be clearly observed that when using our proposed vMF-based trustworthiness metric, the overall performance remains stable, and it achieves the highest average score across all benchmarks. This demonstrates the superiority of the vMF-based trustworthiness metric.

E. Case Study

We have selected more cases for illustration to better demonstrate the effectiveness of DOUBT. Their detection processes are consistent with our initial idea of increasing the diversity of the answer space. Even though a few cases may not show this effect clearly, they can still be detected through the more stable vMF-based trustworthiness metric.

Question: How many uncut fruits are in the image?
GTAns: There are three uncut fruits in the image.
MLLMAns: There are four uncut fruits in the image.
Original Answer Set:
- There are three uncut fruits in the image.
- There are four uncut fruits in the image.
- There are four uncut fruits in the image.
- There are three uncut fruits in the image.
- There are four uncut fruits in the image.
- There are four uncut fruits in the image.
- There are three uncut fruits in the image.
- There are four uncut fruits in the image.
- There are three uncut fruits in the image.
- There are four uncut fruits in the image.
Object List: {Fruit, Fruit, Fruit}
OLB Answer Set:
- 4 - 4 -5 - 4 -3 - 4 - 5 - 4 - 2 - 4

Semantic Entropy: 0.971 ✗
EigenScore: -2.325 ✗
DOUBT: 0.269 ✓

Question: What is the total price for a bottle of Merlot and a bottle of Cabernet shown in the image?
GTAns: 249.98
MLLMAns: 248.98
Original Answer Set:
- 249.98 - $249.98 - 249.98 - 248.98 - 248.98 - 248.98 - 248.98 - 249.98 - 248.98 - 249.98
Object List: {sign, wine bottle, person}
OLB Answer Set:
- The total price for a bottle of Merlot and a bottle of Cabernet is $249.98.
- The total price for a bottle of Merlot and a bottle of Cabernet shown in the image is $248.98.
- $249.98
- 253.98
- The total price for a bottle of Merlot and a bottle of Cabernet shown in the image is 248.98.
- The total price for a bottle of Merlot (139.99) and a bottle of Cabernet (109.99) is 249.98.
- The total price for a bottle of Merlot and a bottle of Cabernet shown in the image is 248.98.
- The total price for a bottle of Merlot and a bottle of Cabernet shown in the image is $249.98.
- 249.98
- The total price for a bottle of Merlot and a bottle of Cabernet shown in the image is 248.98.

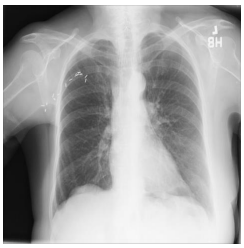
Semantic Entropy: 1.0 ✗
EigenScore: -2.155 ✗
DOUBT: 0.275 ✓

Question: Are which one of these two jets are larger in size?
GTAns: same.
MLLMAns: The jet on the left is larger in size.
Original Answer Set:
- *The jet in the foreground is larger in size.*
- *The jet on the left is larger in size.*
- *The jet on the left is larger in size.*
- *The jet on the right is larger in size.*
- *The jet on the left is larger in size.*
- *The jet in the foreground is larger in size.*
- *The jet in the foreground is larger in size.*
- *The jet on the left is larger in size.*
- *The jet on the left is larger in size.*
- *The jet on the left is larger in size.*

Semantic Entropy: 0.971 ✗
EigenScore: -2.325 ✗
DOUBT: 0.269 ✓

Object List: {Fighter jet, clouds, ocean}
OLB Answer Set:
- *The fighter jet in the foreground appears to be larger than the one in the background.*
- *The jet in the foreground is larger in size.*
- *The jet in the foreground appears to be larger than the one in the background.*
- *The jet in the foreground appears to be larger in size compared to the one in the background.*
- *The first jet is larger in size.*
- *The jet in the foreground is larger in size.*
- *The jet in the foreground is larger in size.*
- *The fighter jet in the foreground is larger in size compared to the one in the background.*
- *The jet on the left is larger in size.*
- *The jet in the foreground is larger in size.*

Question: Which chest xray disease can be found? (Options include: Atelectasis; Cardiomegaly; Effusion; Infiltration; Mass; Nodule; Pneumonia; Pneumothorax; Consolidation; Edema; Emphysema; Fibrosis; Pleural Thickening; Hernia.)
GTAns: Hernia
MLLMAns: Atelectasis
Original Answer Set:
- *Cardiomegaly - Atelectasis - Atelectasis - Cardiomegaly - Atelectasis - Atelectasis - Atelectasis - Atelectasis - Cardiomegaly - Atelectasis*



Semantic Entropy: 0.881 ✗
EigenScore: -2.319 ✗
DOUBT: 0.300 ✓

Object List: {Lung, heart, ribs}
OLB Answer Set:
- *Cardiomegaly - Cardiomegaly - Atelectasis - Atelectasis - Atelectasis - Atelectasis - Cardiomegaly - Atelectasis - Atelectasis - Cardiomegaly*

Question: Is the surfer to the left or to the right of the surfboard?
GTAns: left
MLLMAns: The surfer is to the right of the surfboard
Original Answer Set:
- The surfer is to the right of the surfboard.
- The surfer is to the left of the surfboard.
- The surfer is to the right of the surfboard.
- The surfer is to the right of the surfboard.
- The surfer is to the left of the surfboard.
- The surfer is to the left of the surfboard.
- The surfer is to the right of the surfboard.
- The surfer is to the right of the surfboard.
- To the left.
- The surfer is to the left of the surfboard.

Semantic Entropy: 0.971 ✗
EigenScore: -2.325 ✗
DOUBT: 0.269 ✓

Object List: {surfer, surfboard, ocean}
OLB Answer Set:
- The surfer is to the left of the surfboard.
- The surfer is to the right of the surfboard.
- To the right
- The surfer is to the left of the surfboard.
*- To determine if the surfer is to the left or right of the surfboard, let's analyze the image description provided:\n\n1. **Identify the Surfer**: The image shows a surfer who is in the air, performing a maneuver.\n2. **Identify the Surfboard**: The surfboard is clearly visible and the surfer is standing on it.\n3. **Relative Position**: The surfer is positioned on the surfboard, which means they are directly on top of it.\n\nBased on the image description, the surfer is on top of the surfboard, so they are neither to the left*
- The surfer is to the left of the surfboard.
- The surfer is to the right of the surfboard.
- The surfer is to the right of the surfboard.
- The surfer is to the left of the surfboard.
- The surfer is to the left of the surfboard.

Question: Which continent is highlighted?
GTAns: Asia
MLLMAns: Europe
Original Answer Set:
- Asia - Asia - Europe - Europe - Europe
- Asia - Europe - Europe - Europe - Europe

Semantic Entropy: 0.881 ✗
EigenScore: -2.319 ✗
DOUBT: 0.300 ✓

Object List: {Europe, Asia, North America}
OLB Answer Set:
- Europe and Asia - Europe - Europe - Europe and Asia - Europe - Europe - Europe and Asia - Asia - Europe and Asia -Asia

Question: Is the person bigger than the car?
GTAns: no
MLLMAns: Yes.
Original Answer Set:
- Yes. - Yes. - Yes. - Yes, the person is bigger than the car.
- no - yes - yes - yes - Yes. -Yes.
Object List: {man, bicycle, car}
OLB Answer Set:
- Yes, the person is bigger than the car.
- Yes
- Yes
- Yes, the person is bigger than the car.
- Yes
- Yes, the person is bigger than the car.
- No, the person is not bigger than the car.
- Yes, the person is bigger than the car.
- Yes, the person is bigger than the car.
- Yes, the person is bigger than the car.

Semantic Entropy: 0.922 ✗
EigenScore: -1.956 ✓
DOUBT: 0.465 ✓

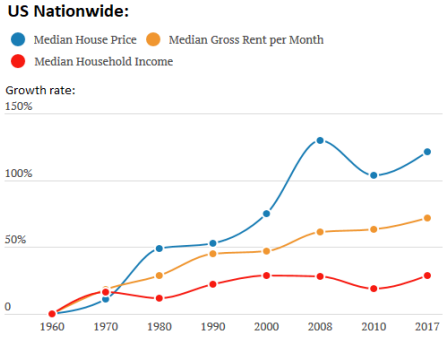
Question: How many tomatoes are there?
GTAns: 5
MLLMAns: There are 6 tomatoes.
Original Answer Set:
- There are six tomatoes.
- There are six tomatoes.
- There are 6 tomatoes.
- There are 6 tomatoes.
- There are six tomatoes.
- There are six tomatoes.
- There are 8 tomatoes.
- There are 7 tomatoes.
- There are six tomatoes.
- There are 6 tomatoes.
Object List: {apple, tomato, jar}
OLB Answer Set:
- There are 6 tomatoes in the image.
- There are 7 tomatoes in the image.
- There are 7 tomatoes in the image.
- There are 9 tomatoes in the image.
- There are 6 tomatoes in the image.
- There are a total of 7 tomatoes in the image.
- There are 6 tomatoes in the image.
- There are six tomatoes in the image.
- There are 6 tomatoes in the image.
- There are six tomatoes in the image.

Semantic Entropy: 0.922 ✗
EigenScore: -2.039 ✓
DOUBT: 0.458 ✓

Question: What is the name of this dish?
GTAns: Caprese Salad
MLLMAns: Caprese Salad
Original Answer Set:
- Caprese Salad - Caprese Salad Skewers - Caprese Salad
- Caprese Salad - Caprese - Caprese Salad. - Caprese skewers.
- Caprese Salad - Caprese Salad - Caprese Salad
Object List: {Tomato, Mozzarella, Plate}
OLB Answer Set:
- Caprese Salad - Caprese Salad - Caprese Salad
- Caprese Salad - Caprese Salad - Caprese Salad
- Caprese Salad - Caprese - Caprese Salad
- Caprese Salad

Semantic Entropy: 0.722 ✓
EigenScore: -1.848 ✗
DOUBT: 0.525 ✓

Question: Which year has the highest growth rate of median house price?
GTAns: 2008
MLLMAns: 2008
Original Answer Set:
- 2007 - 2008 - 2008 - 2008 - 2008
- 2008 - 2009 - 2008 - 2008 - 2008
Object List: {Median House Price, Median Gross Rent per Month, Median Household Income}
OLB Answer Set:
- 2008 - 2008 - 2008 - 2008 - 2008
- 2008 - 2008 - 2008 - 2008 - 2008



Semantic Entropy: 0.922 ✓
EigenScore: -2.072 ✗
DOUBT: 0.660 ✓